Acta Scientifica Balcanica, **2025, 6(1)**
DOI: 10.63356/asb.2025.002 **UDC 599.93+519.17]:004.738.5**
Original scientific paper

THE UNINTENDED CONSEQUENCES OF ALGORITHMIC TRANSPARENCY ON TRUST: A MULTI-DISCIPLINARY ANALYSIS

Dimitrije D. Čvokić[1*]

[1]*University of Banja Luka, Faculty of Natural Sciences and Mathematics, Mladena Stojanovića 2, 78000 Banja Luka, Republic of Srpska, Bosnia and Herzegovina*
***Corresponding author**: dimitrije.cvokic@pmf.unibl.org

Abstract

This paper challenges the widely held assumption that increased algorithmic transparency universally enhances user trust. Through interdisciplinary analysis spanning legal, ethical, and technical domains, we demonstrate that, paradoxically, excessive transparency can erode trust in algorithmic systems. Using agent-based simulations, we model how users with varying cognitive thresholds respond to transparency events, revealing what can be called the "transparency overload" phenomenon. Our findings suggest that strategic opacity—the deliberate limitation of certain types of algorithmic disclosure—may better preserve trust in specific contexts. We propose "context-dependent transparency" as an alternative framework that balances accountability with user' cognitive limitations. This research has significant implications for policymakers and system designers seeking to build genuinely trustworthy algorithmic systems rather than merely transparent ones.

Key words: algorithmic transparency, cognitive overload, strategic opacity, trust preservation, agent-based modeling

INTRODUCTION

In contemporary discourse on algorithmic governance, transparency has emerged as a dominant paradigm and ethical imperative. The mantra that "sunlight is the best disinfectant" (Brandeis, 1914) has been enthusiastically applied to algorithmic systems, with transparency championed as the solution to problems of accountability, fairness, and trust (Ananny and Crawford, 2018). In this paper, we adopt a precise definition of trust: a psychological attitude or relational stance in which an agent accepts vulnerability to another agent, system, or institution based on the belief or expectation that the other will act competently, reliably, and with goodwill within a given context.

Legislative efforts such as the European Union's General Data Protection Regulation (GDPR) and the more recent AI Act enshrine a "right to explanation" and mandate various transparency measures. Industry has responded with "explainable AI" initiatives, transparency reports, and increasingly detailed privacy policies.

Yet emerging evidence suggests a more complex relationship between transparency and trust. We consider the following paradoxes:

1. Privacy notices have proliferated to such an extent that they create “consent fatigue” (Schermer *et al.*, 2014), with one study finding that “reading all privacy policies encountered in a year would take the average person 76 working days” (McDonald and Cranor, 2008).
2. When Google revealed how its flu prediction algorithm worked, “public trust declined rather than increased once limitations were exposed” (Lazer *et al.*, 2014).
3. Research on "algorithmic aversion", termed by Dietvorst *et al.* (2015), demonstrates that “humans often reject algorithms once they observe them making mistakes, even when those algorithms outperform human judgment overall” (Dietvorst *et al.*, 2015).

These examples suggest that the relationship between transparency and trust is non-linear and context-dependent. While inadequate transparency clearly undermines trustworthiness, we propose that excessive transparency can be equally problematic. This paper introduces the concept of "trust-preserving opacity"—the strategic limitation of certain types of algorithmic disclosure to maintain appropriate levels of user trust. This study tests the following hypotheses:

(H1) Algorithmic transparency enhances trust only within specific cognitive thresholds.

(H2) Excessive disclosure leads to trust erosion through cognitive overload.

(H3) Strategically limited disclosure (“strategic opacity”) can ethically preserve warranted trust.

To address these questions, we employ an interdisciplinary approach combining theoretical analysis with agent-based modeling to simulate the dynamics of trust under varying transparency conditions. Cybersecurity research shows similar dynamics, where fusion-based modeling of attack surfaces demonstrates how system disclosure can inadvertently create exploit vectors (Korać *et al.*, 2022). Our findings challenge the assumption that absolute transparency should be the universal goal of algorithmic governance, suggesting instead a more nuanced approach that we term "context-dependent transparency".

LITERATURE AND POLICY REVIEW

THE TRANSPARENCY IMPERATIVE IN AI ETHICS

Transparency has become a cornerstone of ethical AI frameworks worldwide. The European Union's High-Level Expert Group on AI (2019) identifies transparency as one of the seven key requirements for trustworthy AI. Similarly, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) emphasizes transparency as crucial for building public confidence. This emphasis reflects what Ananny and Crawford (2018) call the “transparency imperative” the assumption that seeing inside a system is both necessary and sufficient for accountability. We define transparency as the accessibility and comprehensibility of system processes and decisions to relevant stakeholders, emphasizing effective understanding over maximal disclosure.

The legal operationalization of this imperative is most evident in the GDPR "right to explanation" (GDPR (2016), Wachter *et al.*, 2017) which requires that data subjects receive "meaningful information about the logic involved" in automated decisions (Article 15, GDPR). The EU AI Act (proposed in 2021) goes further, creating tiered transparency requirements based on risk levels, with "high-risk" AI systems subject to extensive documentation and explainability requirements.

However, research in behavioral economics and cognitive psychology suggests limitations to this approach. Tversky and Kahneman's (1974) work on cognitive biases demonstrates that "humans process information using heuristics that can be overwhelmed by excessive detail". More recently, Nissenbaum (2011) has criticized the "transparency paradox" (the term introduced by Nissenbaum), wherein privacy notices simultaneously contain too much information (from a cognitive load perspective) and too little (from a genuine informational perspective).

TRUST EROSION THROUGH OVER-DISCLOSURE

Empirical studies increasingly document cases where transparency initiatives have backfired. In a comprehensive review, Grimmelikhuijsen and Meijer (2014) found that "government transparency often decreased rather than increased citizen trust, particularly when it revealed complexity or internal disagreements". This "transparency trap", as termed by Grimmelikhuijsen and Meijer (2014), operates through several mechanisms:

1. Information Overload: When Facebook increased the granularity of its privacy controls, users became less rather than more engaged with privacy management (Tucker, 2014). The cognitive burden of processing extensive options paradoxically reduced user agency.
2. Uncertainty Amplification: Vaccaro *et al*. (2019) demonstrated that when Instagram began labeling sponsored content more transparently, user trust in all content declined, suggesting that highlighting some problematic aspects can create suspicion about unmarked content.
3. Perfect as Enemy of the Good: Dietvorst *et al.*'s (2015) seminal work on algorithmic aversion shows that once users see algorithms make mistakes—even when those algorithms outperform humans—they lose trust disproportionately, a phenomenon Stray (2021) calls "transparency penalty".
4. Security Vulnerabilities: In technical domains, Pieters (2011) argues that complete transparency about security systems can create new vulnerabilities by providing road maps for attackers, creating a "security through obscurity" dilemma. This challenge is intensified in distributed environments such as IoT and drone networks, where lightweight authentication protocols demonstrate both the necessity and risks of system-level transparency (Bhattarai *et al.*, 2024).

These findings suggest that transparency's relationship to trust follows what behavioral economists call an "inverted U-curve" (Kahneman, 2011)—where both too little and too much transparency diminish trust. The optimal point on this curve likely varies by context, user sophistication, and the nature of the algorithmic system in question.

COGNITIVE LOAD THEORY AND ALGORITHMIC EXPLANATIONS

To understand why excessive transparency can reduce trust, we turn to cognitive load theory (Sweller, 1988). This framework describes how the working memory limitations constrain information processing. When explanations exceed cognitive thresholds, users experience cognitive overload, leading to disengagement, anxiety, or reliance on simplistic heuristics.

Recent work applying cognitive load theory to algorithmic explanations is particularly relevant. Narayanan *et al.* (2018) demonstrated that technical explanations of how machine learning models function often exceed the cognitive capacity of even technically trained users. Miller (2019) found that people tend to understand explanations better when limited to a small number of causes, whereas algorithms often rely on much more complex feature sets. Several empirical studies reinforce these theoretical concerns:

- Wang *et al.* (2020) found that participants exposed to detailed explanations of credit scoring algorithms reported feeling more anxious and less confident in their understanding than those given simplified explanations.
- Dodge *et al.* (2019) showed that “explanations focusing on a few key features were more effective at helping users predict algorithmic behavior than comprehensive feature-importance lists”.

In the line with these studies, we can ask ourselves a natural question: Does excessive disclosure of algorithmic limitations paradoxically reduce users' ability to identify cases where algorithmic advice should be overridden?

These findings suggest that explanation interfaces face an inherent tension between completeness and comprehensibility—a tension that current transparency mandates often fail to acknowledge.

THEORETICAL FRAMEWORK: INTEGRATING LEGAL, ETHICAL, AND TECHNICAL PERSPECTIVES

LEGAL TENSIONS: RIGHT TO EXPLANATION VS. PRACTICAL LIMITATIONS

The legal discourse on algorithmic transparency reveals a fundamental tension between aspirational rights and practical implementation. While the GDPR established a qualified “right to explanation”, legal scholars debate its scope and enforceability (Wachter *et al.*, 2017). More importantly, compliance with transparency regulations has often resulted in practices that satisfy legal requirements without meaningfully empowering users.

For example, Waldman (2018) documents how companies implement "privacy theater"—elaborate consent mechanisms that technically comply with transparency requirements while being designed to discourage actual engagement. Similarly, Buhmann *et al.* (2020) argue that AI ethics principles often function as “ethics washing”, providing rhetorical cover without substantive accountability.

From a legal perspective, the challenge lies in designing transparency requirements that acknowledge cognitive limitations while still ensuring meaningful oversight. The "notice and consent" model of transparency may need to be supplemented or replaced with what

Hildebrandt (2015) calls “transparency enhancing technologies” that make algorithmic impacts visible without requiring users to process complex explanations.

ETHICAL PARADOXES: AUTONOMY VS PRACTICAL UTILITY

From an ethical standpoint, transparency is typically justified as enabling autonomy—people cannot make informed choices about algorithmic systems without understanding how they work. However, this framing assumes that more information always enhances autonomy, an assumption challenged by research on choice architecture (Thaler and Sunstein, 2008).

When transparency mechanisms overwhelm users' cognitive capacity, they can paradoxically reduce meaningful autonomy by forcing choices without understanding. As Nissenbaum (2011) argues, “Transparency, as the public display of information, must be distinguished from transparency, as in 'I see what's going on.'” Only the latter genuinely supports autonomy.

This creates what we term the “transparency-autonomy paradox”: excessive transparency can undermine the very autonomy it aims to support. Resolving this paradox requires acknowledging that autonomy is not simply about the quantity of information but about the quality and accessibility of information relative to cognitive capacity.

TECHNICAL CONSTRAINTS: EXPLAINABILITY TRADE-OFFS

From a technical perspective, the challenge of algorithmic transparency is complicated by fundamental trade-offs between model performance and explainability. As Rudin (2019) argues, many high-performing machine learning models (particularly deep learning approaches) are inherently opaque, creating a tension between accuracy and explicability. Recent computational engineering approaches also propose modeling safety and security boundaries in complex AI-driven systems, highlighting the trade-off between high accuracy and interpretability (Korać *et al.*, 2025a).

Attempts to render such systems transparent typically employ post-hoc explanation methods like Local Interprtable Model-agnostic Explanations (LIME) (Ribeiro *et al.*, 2016) or Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017). However, these methods have significant limitations:

1. They may provide simplified approximations that misrepresent the actual model behavior.
2. They often fail to capture interaction effects between features.
3. They can be manipulated to hide biases while appearing transparent (Aivodji *et al.*, 2019).

The technical reality is that perfect transparency is often unattainable, particularly for complex systems. As Kroll (2018) argues, we may need to shift from the transparency of mechanism to the transparency of process and governance—focusing less on how algorithms work internally and more on how they are developed, validated, and monitored.

This integration of legal, ethical, and technical perspectives suggests a more nuanced approach to transparency—one that acknowledges inherent trade-offs and seeks to optimize rather than maximize disclosure based on context.

AGENT-BASED SIMULATION: MODELING TRUST DYNAMICS UNDER VARYING TRANSPARENCY CONDITIONS

To explore the relationship between transparency and trust more systematically, we developed a simple and illustrative agent-based model simulating user trust responses to varying levels of algorithmic transparency. This approach allows us to test hypotheses about the non-linear relationship between transparency and trust in a controlled environment.

SIMULATION DESIGN

Our simulation was implemented using Mesa, a Python-based modeling framework for agent-based simulation. We created a population of agents representing users with varying cognitive thresholds for processing information about algorithmic systems. These agents interact with a simulated algorithmic system that provides different levels of transparency about its decision-making process.

The key components of our model include:

1. TrustAgent Class: Represents users with varying cognitive capacities and initial trust levels.
2. TransparencyEvent Class: Represents instances where the system provides explanations of varying complexity.
3. AlgorithmicSystem Class: Simulates a decision-making system that interacts with agents.

The central mechanism in our model is how agents update their trust in response to transparency events based on their individual cognitive thresholds. If the complexity of a transparency event exceeds an agent's threshold, trust decreases (representing cognitive overload); if complexity is moderate relative to the threshold, trust increases.

KEY PARAMETERS

Our simulation includes several configurable parameters:

- Agent Count: Number of agents in the simulation (default: 100)
- Cognitive Threshold Distribution: Parameters for the distribution of cognitive thresholds across the agent population (default: normal distribution with $\mu=0.5$, $\sigma=0.15$)
- Initial Trust Distribution: Parameters for the distribution of initial trust levels (default: normal distribution with $\mu=0.5$, $\sigma=0.1$)
- Transparency Regime: Configuration for transparency events (frequency, complexity levels)
- Simulation Duration: Number of time steps to run (default: 100)

These parameters allow us to simulate various scenarios, from high-transparency regimes (frequent, complex explanations) to moderate-transparency regimes (selective, simplified explanations).

Table 1. Mapping between theoretical constructs and simulation parameters

Theoretical Construct	Simulation Parameter	Expected Outcome
Cognitive Load Theory (Sweller, 1988): Information beyond cognitive threshold causes overload	cognitive_threshold variable in each agent	Agents with lower thresholds show faster trust decline when exposed to high-complexity transparency events
Transparency Regime (high vs. moderate)	transparency_regime parameter ("high" vs. "moderate")	High transparency → trust erosion; moderate transparency → sustained trust
Trust Overload Phenomenon ("transparency penalty")	trust_change function based on overload count	Accumulated overload amplifies negative trust adjustment
Recovery Dynamics (trust hysteresis)	overload_count memory in agent behavior	Trust restoration requires multiple positive interactions

RESULTS AND ANALYSIS

We ran multiple simulations comparing high-transparency regimes (transparency levels consistently exceeding the cognitive threshold of approximately 70% of agents) with moderate-transparency regimes (transparency levels exceeding the threshold for only 30% of agents).

Our findings revealed several key patterns:

1. Trust Collapse in High-Transparency Regimes: Under conditions of high transparency, overall trust initially increased but then declined significantly over time, with approximately 60% of agents eventually developing trust scores below their initial levels.
2. Sustained Trust in Moderate-Transparency Regimes: With moderate transparency, trust increased more gradually but was sustained over time, with 80% of agents maintaining trust scores above their initial levels throughout the simulation.
3. Threshold Effects: Agents with low cognitive thresholds (bottom quartile) showed 40% faster trust decline when exposed to high-transparency regimes compared to the general population.
4. Recovery Dynamics: Once trust declined due to transparency overload, it required approximately twice as many positive interactions to restore trust to previous levels, suggesting a hysteresis effect in trust dynamics.

Figure 1 visualizes these results, showing trust trajectories under high versus moderate transparency regimes. The simulation demonstrates how excessive transparency can trigger what we term a "trust collapse cascade"—where initial instances of cognitive overload spread through a population as declining trust makes users more susceptible to further overload.

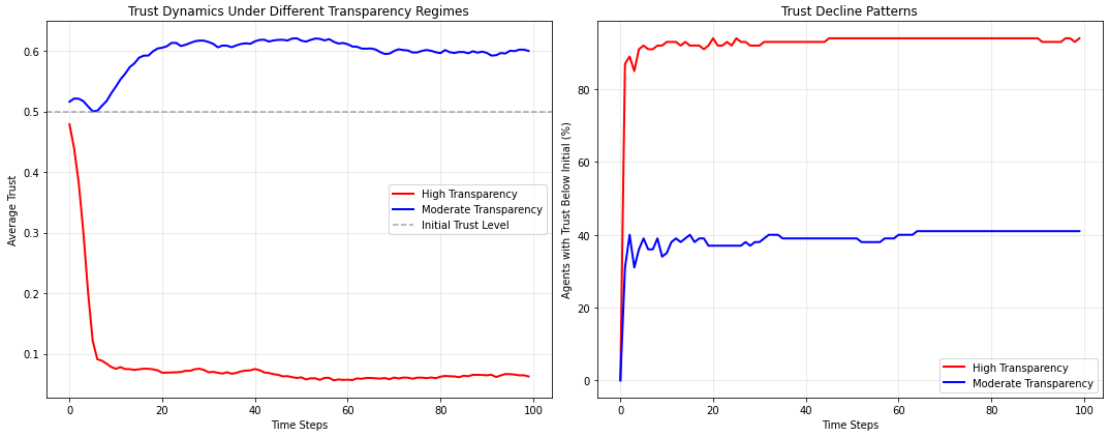

Figure 1. Results demonstrate the "transparency overload", a phenomenon where excessive transparency can paradoxically reduce trust through cognitive overload

LIMITATIONS AND EXTENSIONS

While our simulation provides valuable insights into the dynamics of transparency and trust, it has several limitations that warrant consideration:

1. This model simplifies cognitive processing to a single threshold parameter, whereas the real cognitive responses to explanations are multidimensional.
2. The simulation does not account for social learning or network effects, where users' trust might be influenced by others' experiences.
3. This model treats transparency as uniformly distributed across the agent population, whereas real-world transparency initiatives might be targeted based on user sophistication.

Future extensions could address these limitations by incorporating more sophisticated cognitive models, social network structures, and targeted transparency approaches.

ETHICAL PARADOXES IN ALGORITHMIC TRANSPARENCY

MORAL LUCK AND UNINTENDED CONSEQUENCES

The potential for transparency to reduce rather than enhance trust raises significant ethical questions. Drawing on Thomas Nagel's (1979) concept of "moral luck", we suggest that transparency advocates face a moral dilemma: their well-intentioned efforts may produce outcomes contrary to their goals due to factors beyond their control (specifically, human cognitive limitations).

This creates what philosopher Bernard Williams (1981) calls a "moral residue"—even the right action (increasing transparency) can lead to harm (reduced trust) through no fault of the actor. This perspective challenges the deontological framing often applied to transparency, which treats disclosure as an inherent good regardless of consequences.

PATERNALISM VS. STRATEGIC OPACITY

Strategic opacity—the deliberate limitation of certain types of disclosure—raises concerns about paternalism. Strategic opacity is defined here as the deliberate, ethically motivated limitation of algorithmic disclosure intended to preserve user trust and cognitive

stability without undermining accountability. Unlike mere secrecy or information withholding, strategic opacity is grounded in a principle of means paternalism (Sunstein, 2014), where information is structured to serve users' genuine understanding rather than overwhelm them. Conceptually, it builds on Nissenbaum's (2011) idea of contextual integrity and Rudin's (2019) argument that excessive explainability can distort comprehension. Thus, strategic opacity is not a retreat from transparency, but a calibrated design choice aimed at achieving "effective transparency" rather than "absolute transparency." Critics might argue that withholding information, even to prevent cognitive overload, constitutes a form of paternalism that violates user autonomy.

However, drawing on Sunstein's (2014) concept of "means paternalism", we argue that strategic opacity need not interfere with users' ends (making informed decisions) but merely adjusts the means through which information is provided. This shifts the ethical question from whether opacity is ever justified to how opacity should be implemented to best serve users' genuine interests in understanding algorithmic systems.

TRUST VS. TRUSTWORTHINESS

A crucial ethical distinction in this domain is between trust (a psychological state) and trustworthiness (a normative quality of systems). Trustworthiness, as used here, refers to the objectively verifiable qualities of a system's reliability, integrity, and ethical alignment, regardless of whether it is currently trusted. O'Neill (2002) argues that transparency mechanisms often focus on generating trust without ensuring trustworthiness—creating what she calls "trust surrogates" that provide psychological reassurance without substantive accountability.

Strategic opacity risks exacerbating this problem if it conceals genuine issues with algorithmic systems. Therefore, we argue that strategic opacity is ethically justified only when:

1. It preserves trust in systems that are independently trustworthy.
2. It is implemented alongside robust but less visible accountability mechanisms.
3. Users retain access to more detailed information upon request.

This approach acknowledges that trust has instrumental value only when aligned with trustworthiness, and that transparency serves ethical ends only when it genuinely empowers rather than overwhelms users.

POLICY IMPLICATIONS: TOWARD CONTEXT－DEPENDENT TRANSPARENCY

Context-dependent transparency is defined here as the adaptive calibration of algorithmic disclosure that optimizes user understanding and trust by aligning the amount, form, and timing of transparency with (i) user characteristics, (ii) algorithmic complexity, and (iii) decision context. It extends the notions of selective disclosure (which focuses on what is shown), tiered transparency (how information is layered), and effective transparency (how

well it is understood) by integrating all three dimensions into a unified, context-sensitive framework.

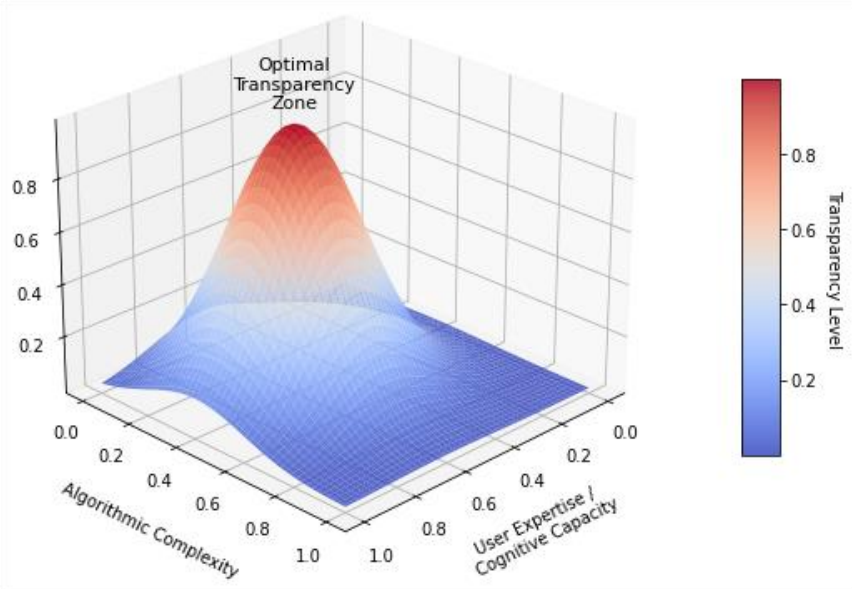

Figure 2. Results demonstrate the "transparency overload" phenomenon where excessive transparency can paradoxically reduce trust through cognitive overload

Our research suggests the need for a paradigm shift in algorithmic governance—moving from what we call "absolute transparency" to what can be named as "context-dependent transparency". This approach recognizes that optimal transparency levels vary based on user characteristics, algorithmic complexity, and decision contexts.

While effective transparency emphasizes the quality and comprehensibility of disclosure, context-dependent transparency extends this idea by situating transparency levels within specific user, system, and risk contexts. In other words, effective transparency is about how information is communicated, whereas context-dependent transparency concerns when, to whom, and to what extent it should be disclosed.

TIERED DISCLOSURE MODELS

One practical application of context-dependent transparency is a tiered disclosure model. Rather than providing all users with identical explanations, systems could offer (also presented in Figure 2):

- Layer 1: Simplified explanations focusing on 2-3 key factors affecting the decision, designed for accessibility.
- Layer 2: More detailed explanations available upon request, including more comprehensive feature importance information.
- Layer 3: Technical documentation for experts, researchers, and regulators, including model specifications and performance metrics.

This approach aligns with cognitive load research suggesting that users process information more effectively when they can control its complexity. It also addresses legal

requirements for explanation while acknowledging practical limitations on human information processing.

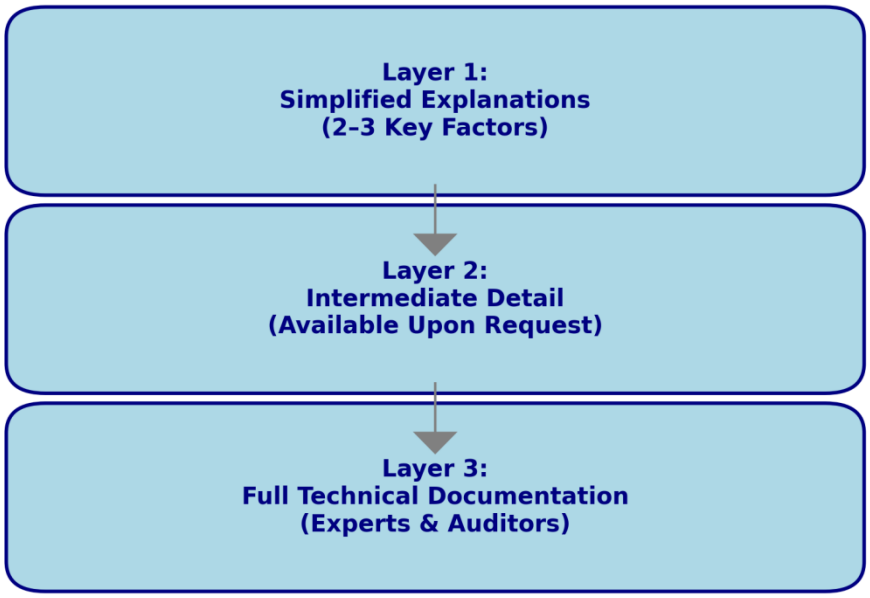

Figure 3. Schematic representation of the tiered disclosure model. Users can navigate between layers on their expertise and information needs

AGGREGATE ACCOUNTABILITY MECHANISMS

Beyond individualized explanations, context-dependent transparency emphasizes the importance of aggregate accountability mechanisms. These include:

- Algorithmic Impact Assessments (AIAs): Pre-deployment evaluations of potential algorithmic harms, published in accessible formats.
- Outcome Monitoring: Ongoing analysis of algorithmic outputs to detect patterns of bias or error.
- Independent Auditing: Third-party verification of algorithmic systems without necessarily disclosing proprietary details.

It is worth to note that AI-based fusion frameworks for metric evaluation can complement AIAs by providing quantitative authentication and evaluation metrics for algorithmic transparency initiatives (Korać *et al.*, 2025b).These mechanisms can ensure accountability without requiring individual users to process complex technical information—shifting the transparency burden from users to governance structures.

USER-CENTERED DESIGN OF EXPLANATIONS

Our findings highlight the importance of designing explanations around user needs rather than technical or legal requirements. This includes:

- Calibrating explanation complexity to user expertise.
- Focusing on counterfactual explanations (how outcomes could be changed) rather than feature attribution alone.
- Using visual and interactive explanations where appropriate.

- Testing explanations empirically for comprehension and utility.

Such user-centered approaches can ensure that transparency initiatives genuinely empower users rather than overwhelming them with information they cannot effectively process.

REGULATORY IMPLICATIONS

For policymakers, our research suggests several refinements to current regulatory approaches:

1. Shift from mandating specific forms of disclosure to requiring evidence-based transparency—demonstrating that explanations actually enhance user understanding.
2. Consider differential transparency requirements based on algorithmic risk and complexity rather than one-size-fits-all mandates.
3. Prioritize "effective transparency" over "absolute transparency", focusing on outcomes (genuine understanding) rather than outputs (quantity of information disclosed).
4. Support research on cognitive aspects of algorithmic explanations to inform evidence-based transparency standards.

These policy directions would help address what Veale *et al*. (2018) call the "explanation gap"—the divergence between technical capabilities, legal requirements, and genuine user needs in algorithmic accountability.

Table 2. Mapping proposed reforms to regulatory provisions

Proposed Reform	Related Regulatory Provision	Alignment/Extension
Evidence-based transparency (empirical verification of user understanding)	GDPR Art. 15 ("Right to explanation"); EU AI Act Art. 13 ("Transparency obligations")	Extends from procedural to outcome-based compliance
Tiered disclosure model (layered explanation levels)	GDPR Recital 58; EU AI Act Annex IV ("Documentation requirements")	Operationalizes "meaningful information" through multi-level design
Cognitive load–aware interface design	EU AI Act Recital 47 (human oversight)	Incorporates cognitive constraints into user interface obligations
Aggregate accountability (independent auditing)	U.S. Algorithmic Accountability Act (2022, Sec. 4)	Complements U.S. audit-based approach with contextual adaptation

DISCUSSION AND LIMITATIONS

ADDRESSING THE "SUNLIGHT AS DISINFECTANT" CRITIQUE

A potential criticism of our argument is that it undermines the well-established principle that "sunlight is the best disinfectant" (Brandeis, 1914). This principle has been central to progressive governance reforms for over a century, and suggesting limitations to transparency may seem regressive.

However, we argue that our approach refines rather than rejects this principle. The metaphor of "sunlight" deserves closer examination—while moderate sunlight enables vision,

excessive sunlight causes blindness. Similarly, transparency enables accountability only when calibrated to human cognitive capacities.

More importantly, we distinguish between transparency for oversight (by regulators, researchers, and civil society organizations) and transparency for individual decision-making. Our argument for strategic opacity applies primarily to the latter, while we continue to advocate for robust transparency for oversight purposes.

THE RISK OF "OPACITY LAUNDERING"

A legitimate concern is that arguments for strategic opacity could be misappropriated to justify unwarranted secrecy—what was termed as "opacity laundering" by Stray (2021). The term opacity laundering refers to the rhetorical or institutional practice of invoking complexity, confidentiality, or cognitive limitations as a pretext for concealing unethical or unaccountable algorithmic behavior. As described by Stray (2021), opacity laundering allows organizations to appear ethically responsible while in fact shielding problematic decision processes from scrutiny. In this paper, we explicitly distinguish such misuse from ethically justified strategic opacity by linking the latter to independently verifiable trustworthiness and robust oversight mechanisms. Companies might claim to be preventing cognitive overload when actually concealing problematic algorithmic practices.

We acknowledge this risk and emphasize that strategic opacity is justified only within the ethical framework outlined in section "Ethical Paradoxes in Algorithmic Trasnparency"—specifically, when it preserves trust in independently trustworthy systems. This requires robust verification mechanisms beyond individual transparency, including the aggregate accountability approaches described in subection "Aggregate Accountability Mechanisms" (previous section).

EMPIRICAL VALIDATION BEYOND SIMULATION

While our agent-based model provides theoretical insights into transparency-trust dynamics, empirical validation in real-world contexts is essential. Future research should test our hypotheses through controlled experiments comparing user trust and understanding under different transparency regimes.

Such research should be longitudinal, as our simulation suggests that the negative effects of excessive transparency may emerge only over time, after repeated exposure to overwhelming explanations. Cross-cultural studies would also be valuable, as cognitive responses to algorithmic explanations likely vary across cultural contexts. Validation frameworks such as the Fishbone Model for authentication systems could be adapted to evaluate transparency–trust dynamics across diverse contexts (Korać and Simić, 2019).

CONCLUSION AND FUTURE DIRECTIONS

This paper has challenged the assumption that algorithmic transparency universally enhances trust, demonstrating through interdisciplinary analysis and computational modeling that excessive transparency can paradoxically erode trust through cognitive overload. We

have proposed "context-dependent transparency" as an alternative framework that balances accountability needs with user cognitive limitations.

Our findings have significant implications for how we conceptualize and implement algorithmic governance. Rather than pursuing absolute transparency as an unqualified good, we suggest calibrating transparency to context—providing different levels of disclosure for different users and purposes while ensuring robust accountability through aggregate mechanisms. Future research directions can be grouped into three categories:

1. Empirical: Comparative user studies and longitudinal analysis of transparency–trust dynamics.
2. Theoretical: Development of adaptive cognitive models and game-theoretic trust propagation frameworks.
3. Regulatory and policy-oriented: Legal analysis of context-dependent transparency and standardization within GDPR/AI Act frameworks.

As algorithmic systems become increasingly integrated into consequential decisions, developing effective rather than merely extensive transparency becomes crucial. Strategic opacity—when ethically implemented—may be essential to preserving the trust necessary for algorithmic systems to serve human ends.

ACKNOWLEDGMENETS

This research was supported by the Government of the Republic of Srpska through a national research grand (No. 1259114). The author gratefully acknowledge this support, which made the completion of this work possible.

REFERENCES

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S. & Tapp, A. (2019). Fairwashing: The risk of rationalization. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California: PMLR 97, 2019. (pp. 161-170).

Ananny, M. & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973-989. https://doi.org/10.1177/1461444816676645

Bhattarai, I., Pu, C., Choo, K.-K. R., & Korać, D. (2024). A lightweight and anonymous application-aware authentication and key agreement protocol for the Internet of Drones. *IEEE Internet of Things Journal*, *11*(11), 19790–19803. https://doi.org/10.1109/JIOT.2024.3367799

Brandeis, L. D. (1914). *Other people's money and how the bankers use it.* Frederick A. Stokes Company.

Buhmann, A., Paßmann, J. & Fieseler, C. (2020). Managing algorithmic accountability: Balancing reputational concerns, engagement strategies, and the potential of rational discourse. *Journal of Business Ethics*, *163*(2), 265-280. https://doi.org/10.1007/s10551-019-04226-4

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114-126. https://doi.org/10.1037/xge0000033

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. In N. Oliver & J. Smith (Eds.), *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, Marina del Ray, CA, USA, March 16–20, 2019 (pp. 275-285). ACM. https://doi.org/10.1145/3301275.3302310

European Commission, High-Level Expert Group on Artificial Intelligence. (2019, April 8). Ethics guidelines for trustworthy AI. Publications Office of the European Union. Retrieved from: https://ec.europa.eu/digital-strategy/en/library/ethics-guidelines-trustworthy-ai

GDPR (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Grimmelikhuijsen, S. G. & Meijer, A. J. (2014). Effects of transparency on the perceived trustworthiness of a government organization: Evidence from an online experiment. *Journal of Public Administration Research and Theory*, *24*(1), 137-157. https://doi.org/10.1093/jopart/mus048

Hildebrandt, M. (2015). *Smart technologies and the end(s) of law: Novel entanglements of law and technology*. Edward Elgar Publishing. ISBN: 9781786430229.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan. ISBN: 9780374533557.

Korać, D., Čvokić, D. & Simić, D. (2025a). Computational engineering approach-based modeling of safety and security boundaries: A review, novel model, and comparison. *Archives of Computational Methods in Engineering*. Advance online publication. https://doi.org/10.1007/s11831-025-10352-2

Korać, D., Damjanović, B., Simić, D. & Choo, K.-K. R. (2022). A hybrid XSS attack (HYXSSA) based on fusion approach: Challenges, threats and implications in cybersecurity. *Journal of King Saud University - Computer and Information Sciences*, *34*(10, Part B), 9284–9300. https://doi.org/10.1016/j.jksuci.2022.02.015

Korać, D., Damjanović, B., Simić, D. & Pu, C. (2025b). Management of evaluation processes and creation of authentication metrics: Artificial intelligence-based fusion framework. *Information Processing & Management*, *62*(6), 104233. https://doi.org/10.1016/j.ipm.2025.104233

Korać, D. & Simić, D. (2019). Fishbone model and universal authentication framework for evaluation of multifactor authentication in mobile environment. *Computers & Security*, *85*, 313–332. https://doi.org/10.1016/j.cose.2019.05.012

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180084. https://doi.org/10.1098/rsta.2018.0084

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*(6176), 1203-1205. https://doi.org/10.1126/science.1248506

Lundberg, S. M. & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*. https://doi.org/10.5555/3295222.3295230

McDonald, A. M. & Cranor, L. F. (2008). The cost of reading privacy policies. *ISJLP, 4*, 543. Retrieved from: https://kb.osu.edu/server/api/core/bitstreams/a9510be5-b51e-526d-aea3-8e9636bc00cd/content

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Nagel, T. (1979). *Mortal questions* (pp. 24-38). Cambridge University Press. Retrieved from: https://rintintin.colorado.edu/~vancecd/phil1100/Nagel1.pdf

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S. & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*. https://doi.org/10.48550/arXiv.1802.00682

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32-48. https://doi.org/10.1162/DAED_a_00113

O'Neill, O. (2002). *A question of trust: The BBC Reith Lectures 2002*. Cambridge University Press. ISBN: 9780521529969.

Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology*, *13*(1), 53-64. https://doi.org/10.1007/s10676-010-9253-3

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Schermer, B. W., Custers, B. & van der Hof, S. (2014). The crisis of consent: How stronger legal protection may lead to weaker consent in data protection. *Ethics and Information Technology*, *16*(2), 171-182. https://doi.org/10.1007/s10676-014-9343-8

Sunstein, C. R. (2014). *Why nudge?: The politics of libertarian paternalism*. Yale University Press. Retrieved from: https://yalebooks.yale.edu/book/9780300212693/why-nudge/

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/10.1016/0364-0213(88)90023-7

Stray, J. (2021). *Show me the algorithm: Transparency in recommendation systems*. Schwartz Reisman Institute for Technology and Society. Retrieved from: https://srinstitute.utoronto.ca/news/recommendation-systems-transparency

Thaler, R. H. & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press. Retrieved from: https://psycnet.apa.org/record/2008-03730-000

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (First ed.). IEEE Standards Association. Retrieved from: https://www.ethics.org/wp-content/uploads/Ethically-Aligned-Design-May-2019.pdf

Tucker, C. E. (2014). Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research, 51*(5), 546-562. https://doi.org/10.1509/jmr.10.0355

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. DOI: 10.1126/science.185.4157.1124

Vaccaro, K., Sandvig, C. & Karahalios, K. (2019). "At the end of the day Facebook does what it wants": How users experience contesting algorithmic content moderation. *Proc. ACM Hum.-Comput. Interact. 4*, CSCW2, Article 167 (October 2020), 22 pages. https://doi.org/10.1145/3415238

Veale, M., Binns, R. & Edwards, L. (2018). Algorithms that remember: Model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180083. https://doi.org/10.1098/rsta.2018.0083

Wachter, S., Mittelstadt, B. & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76-99. https://doi.org/10.1093/idpl/ipx005

Waldman, A. E. (2018). Privacy, notice, and design. *Stanford Technology Law Review*, 21, 74. https://dx.doi.org/10.2139/ssrn.2780305

Wang, R., Harper, F. M. & Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In R. Bernhaupt et al. (Eds), *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, April 25-30, 2020 (pp. 1-14). ACM https://doi.org/10.1145/3313831.3376813

Williams, B. (1981). *Moral luck: Philosophical papers 1973-1980*. Cambridge University Press. https://doi.org/10.1017/CBO9781139165860

APPENDIX 1: SIMULATION CODE

The complete Python-based simulation ("Transparency–Trust Model") has been made available as a runnable script. It can be accessed and downloaded from the following repository:

https://drive.google.com/drive/folders/1Z53UPlyG8pC9lNRVLt-fhRF3HObOcWd7?usp=drive_link

The script includes detailed comments and configuration options corresponding to Table 1 parameters. Researchers can modify transparency regimes, cognitive threshold distributions, and agent behaviors to replicate or extend our results.

APPENDIX 2: POLICY SUMMARY FOR REGULATORS

Executive Summary: Moving from Absolute to Context-Dependent Transparency

Based on our interdisciplinary analysis and computational modeling, we recommend that regulators adopt a "context-dependent transparency" framework for algorithmic governance. This approach recognizes that optimal transparency levels vary based on user characteristics, algorithmic complexity, and decision contexts.

Key Policy Recommendations:

1. Evidence-Based Transparency Standards: Replace mandates for specific disclosure formats with requirements to demonstrate that explanations enhance user understanding through empirical testing.
2. Tiered Disclosure Requirements: Implement multi-layered explanation systems that provide simplified explanations by default, with more detailed information available upon request.
3. Cognitive Load Considerations: Require algorithmic explanation interfaces to be designed with human cognitive limitations in mind, incorporating insights from behavioral psychology.
4. Aggregate Accountability Mechanisms: Supplement individual explanations with systemic oversight tools including AIAs, outcome monitoring, and independent auditing.
5. User-Centered Design Standards: Mandate that explanations be tested for comprehension and utility rather than simply technical completeness.

Implementation Pathway:

- Phase 1 (6 months): Pilot tiered disclosure systems with volunteer organizations
- Phase 2 (12 months): Develop cognitive load guidelines for explanation interfaces
- Phase 3 (18 months): Implement reformed transparency requirements for high-risk AI systems
- Phase 4 (24 months): Establish independent auditing frameworks for algorithmic accountability

Expected Outcomes:

This approach should increase genuine user understanding while reducing cognitive burden, ultimately creating more effective algorithmic accountability than current maximum-disclosure approaches. The framework preserves the accountability benefits of transparency while acknowledging practical limitations on human information processing.

НЕЖЕЉЕНЕ ПОСЉЕДИЦЕ АЛГОРИТАМСКЕ ТРАНСПАРЕНТНОСТИ ПО ПОВЈЕРЕЊЕ: МУЛТИДИСЦИПЛИНАРНА АНАЛИЗА

Димитрије Д. Чвокић[1*]

[1]Универзитет у Бањој Луци, Природно-математички факултет, Младена Стојановића 2, 78000 Бања Лука, Република Српска, Босна и Херцеговина

***Аутор за кореспонденцију**: dimitrije.cvokic@pmf.unibl.org

Сажетак

У овом раду доводи се у питање широко заступљено увјерење да повећана алгоритамска транспарентност универзално појачава корисничко повјерење. Преко мултидисциплинарне анализе—обухватајући правну, етичку, и техничку сферу—показујемо да прекомјерна транспарентност може парадоксално да поткопа повјерење у алгоритамске системе. Кроз симулацију агентног модела истражујемо како корисници с различитим котнитивним праговима реагују на "транспарентност догађаја", откривајући појаву коју називамо намјерно ограничење одређених нивоа објелодањивања алгоритама, те како се може боље очувати повјерење у специфичим контекстима. Предлажемо радни оквир "контекстно-зависна транспарентност" као алтернативу која балансира одговорност са когнитивним бранама корисника. Ово истраживање има значајне импликације за креаторе политика, с фокусом на развијање алгоритамских система који су заиста вриједни повјерења, а не само транспарентности.

Кључне ријечи: алгоритамска транспарентност, когнитивно преоптерећење, стратегијски опацитет, очување повјерења, агентно моделовање

Received July 1, 2025
Accepted October 12, 2025